EXPLOITING TOPIC PRAGMATICS FOR NEW EVENT DETECTION IN TDT-2004

Ronald K. Braun and Ryan Kaneshiro

Stottler Henke Associates, Inc.
951 Mariner's Island Blvd., Suite #360
San Mateo, CA 94404

ABSTRACT

TDT-2004 marked Stottler Henke's second year of participation in the New Event Detection (NED) track of the Topic Detection and Tracking (TDT) evaluations. Our official entry this year consisted of three "pragmatics-based" classifiers operating in a majority voting framework. The system performed well, achieving by small margins the best optimized topic- and story-weighted CFSD scores for participating NED systems. We again validated the hypothesis that ensemble collections of classifiers can outperform the individual classifiers that compose them. Performance over the new TDT5 corpus was worse relative to previous corpora and overall accuracy within the NED community remains significantly below operationally desirable levels. We present a brief summary of our second year approach and a preliminary characterization of our performance results based on the experimental runs submitted to the TDT-2004 evaluation.

1. INTRODUCTION

Stottler Henke has developed a workbench application called TOPIC ("Topic-Oriented Pragmatics and Invariant Chaining") for prototyping and evaluating different pragmatics-based techniques in event detection and topic tracking.* We postulate the existence of a variety of pragmatic processes and features that structure a news story as it unfolds over time. For each such feature that can be made computationally accessible, we implement a classifier that attempts the NED task using that feature as its basis for topic novelty judgment. These classifiers are housed in a committee architecture that applies an evidence combination technique to synthesize a global view of story novelty. Because an ensemble view of novelty is generated, no particular classifier need operate with perfect accuracy.

A fundamental premise underlying our work in pragmatics-based new event detection is that multiple structuring processes operate throughout the evolution of a story from the occurrence of events in the world to the reporting of those events to the consumption of resulting news stories by a target audience. These structuring processes may contain information that cues for story novelty and might thus be exploited by a NED system. We loosely define pragmatics as "non-semantic structure arising from *how* a topic is reported through time." With this focus we mean to avoid formal semantic modeling and a reliance on purely statistical linguistic techniques in an effort to bring to light other structuring aspects of news story text.

We have developed a basic Pragmatics Framework that we use as an idea pump for generating classifiers. This framework identifies possible structuring processes that arise from stages in the evolution from incipient event to final consumer product. These stages and associated pragmatic structuring include: the occurrence of some triggering event within the world, wherein events reported in a news story occur in the natural world and as such are subject to physical laws that structure them; event reporting, wherein a triggering event is observed by reporting agents who then summarize and contextualize the event within news reports; language expression, wherein news reports are encoded within a particular language and possess grammatical and co-occurrence regularities; media presentation, wherein news is conveyed in stereotypical formats defined by the media of presentation; and audience consumption, wherein information is filtered and customized with respect to the interests and intents of its audience. This Pragmatics Framework is detailed in [2] and so we omit further characterization here.

Each developed classifier exists to make a new-event detection decision with respect to every incoming story. Some means of combining the individual outputs of all classifiers is necessary to determine the final system NED verdict. Rather than training a single classifier over a range of features derived from the Pragmatics Framework, we have opted to house individual, one-feature classifiers in a committee-based architecture, utilizing majority voting or ensemble learning techniques to combine their results. New classifiers can be installed or removed at will and their contributions to the final system judgment evaluated. Various evidence combination techniques include classifier-independent techniques like majority voting schemes, which operate without specific knowledge of the individual constituents contained by the committee, and classifier-aware methods, including Bayesian and regression techniques, which attempt to learn classifier weightings based on particular committee configurations of classifiers with known properties. These have also been described at length in [2].

The remainder of this paper will summarize our participation in TDT-2004. Section 2 discusses several error classes that motivated some of our classifier development and reviews the

* This work is supported through DARPA SBIR contract DAAH01-03-C-R108.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2004		**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Exploiting Topic Pragmatics for New Event Detection in TDT-2004	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Defense Advanced Research Projects Agency,3701 North Fairfax Drive,Arlington,VA,22203-1714**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**8**	

prospects for significant new progress in NED. Section 3 examines the impact of the new TDT5 corpus on our system. Section 4 describes the classifiers used and Section 5 the evidence combination techniques employed in the five evaluation runs submitted by Stottler Henke. Section 6 concludes with a summary of our evaluation results.

2. ERROR ANALYSIS

To better understand the limitations of our current operational system, we conducted an informal error analysis of one configuration of TOPIC over the TDT3 newswire corpus. We ran a pipelined committee consisting of the baseline full-text classifier followed by a sentence linking classifier (achieving topic-weighted CFSD = 0.4912, with pMiss = 0.4000 and pFA = 0.0186). We non-randomly selected 50 erroneous judgments made by the system (30 misses and 20 false alarms, selected to try and maximize the breadth of error classes covered) and evaluated whether the judgment was in fact incorrect and, if so, whether a computational system could reasonably be expected to make a correct judgment and what would be necessary for it to do so. We estimated that approximately 28% of misses and 35% of the false alarms in our sample had computationally visible features that made them potentially tractable. We believe the remaining errors cannot be corrected because they belong to error classes that are simply not addressable by NED systems under current evaluation conditions (these classes are detailed shortly).

Plugging these reduced error rates into the CFSD computation (given in [4]), and with caveats for the small sample size, various topic and error independence assumptions, the informality of the experiment, and the back-of-the-envelop calculations used, we conclude that an operational ballpark CFSD limit of 0.35 exists for NED systems operating on the TDT3 corpus. This suggests that current NED systems are actually performing fairly close to a limit on accuracy imposed by the evaluation conditions. Additionally, we believe the current assumption of one topic per story biases the evaluation to favor approaches that employ full-text techniques, since such techniques implicitly encode the one topic assumption, where the topic is taken to be the story *in toto*. Unfortunately, several plausible techniques that might be used to chip away at tractable error classes rely on sub-story-level granularity analysis. Techniques that decompose stories into constituent events, for example, may be penalized because they detect additional subtopics. We are therefore somewhat pessimistic that dramatic improvements to existing systems can be made within the current evaluation paradigm. This said, we do feel that significant interesting questions remain to be tackled within the NED track, many of them deriving from attempts to remedy the error classes detailed next.

Recall that a *miss* type error is one in which a NED system says that a story is not a first story even though it in fact is the first for an evaluation topic. Alternately, a *false alarm* type error is one in which a system says that a story is a first story when it is in fact not the first for an evaluation topic. Several classes of problem conspire to confound NED systems into making these errors.†

Annotation effects. Topic annotations are performed by the LDC under time constrained circumstances [3], so it is to be expected that on-topic stories may have been missed and so are mislabeled in the topic evaluation tables. (I don't believe we've ever found a story marked as on-topic that we would dispute, however.) We discovered two instances of missed stories in our examination of the 50 errors described previously. These may have been an artifact of the keyword search process used by the LDC, as the stories were describing the same topical events but one lacked an obviously useful retrieval keyword found in the other story. Such fragmentation, where one story is elaborated by another that makes explicit the topical relevance of the two stories, is not entirely uncommon.

Lack of a priori topic definitions. The topics used for evaluation are structured by the LDC's topic definitions; that is, the topicality of a story is usually constrained by certain criteria presented to the annotators during corpus annotation. These criteria are simply not visible to computational systems performing topic tracking and so it should come as no surprise that finer distinctions in story topicality cannot be made by NED systems. In performing our review of the 50 errors, for example, in most cases it was quite obvious that the stories were "on topic" in some sense; we frequently had to consult the topic definitions to determine exactly what criteria were being applied to the topic at hand to determine why one was on-topic and the other off-topic. For example, two stories described in nearly identical terms different accidents that occurred in Kiev, Ukraine within a one week period as a result lackluster maintenance due to declining economic conditions. One story occurred in a factory, one in a mine. The topic description specified that only mining accidents were on topic. Without this *a priori* structuring knowledge, errors of this sort are unavoidable. The problem of seed sensitivity, wherein topic definitions are sensitive to a founding seed story despite other plausible levels of generality or specificity in topic definition, exacerbates this error class.

Lack of semantic knowledge. In many cases, semantic or domain knowledge must be brought to bear to recognize that two instances of a general phenomena are linked topically. For example, details of two election races for different individuals may be topically related if they are both running for the same class of position (e.g., midterm congressional races). Cause and effect relationships that semantically relate two stories are frequently not made explicit within causally related stories. General topic knowledge must also be codified to some degree. For example, a topic is required to follow from a triggering seed event, so two instances of the same type of meeting by the same organization are seldom on-topic if they do not share the same spatiotemporal context.

Multiple topics within a story. A significant problem arises when multiple topics compose the first story for an evaluation topic. If any of those non-official topics match a previous story, a miss

† There is another error metric that is not detected by the NED evaluation conditions, which we call the *lucky* error rate. A lucky error occurs when a system correctly judges a story to be non-first, but does so for incorrect reasons (e.g., matching a story to a non-topical story). This rate runs at around 12.25% for our baseline methodology. The sources for these errors are likely to be identical to those underlying misses in general.

error will result with respect to the evaluation topic. Alternately, if the story events are quite different in character than the main topic event (say because the story is from a new activity of an evolving topic), the linking references to the topic event may be swamped by the text of the new activity, generating a false alarm. This phenomenon is prevalent for preparatory or preliminary events that occur prior to a future main topic seed event. Those events tend to be discussed in localized story clusters (e.g., the training and selection of an Olympic team in South Korea), with the future event (e.g., the Asian Games) only being alluded to with a linking sentence. Additionally, related or previous events may be mentioned via sentence links to contextualize or analogize events in a story; some of these referenced topics are only exemplars, while others further develop the reader's understanding of the primary topic.

High overlap of entities due to subject marginality or class membership. Many stories about small countries outside of the target audience's interest areas (e.g., Lebanon or the Koreas in the TDT3 corpus) all reference virtually identical people and places (e.g., the president of country, the capitol city), regardless of the events of the story, confounding NED systems. Extreme care must be taken to ensure that at least one spatiotemporal event overlaps in these cases. Similarly, multiple instances of the same type of event (e.g., types of natural disasters, meetings of specific organizations) all share common vocabulary and co-occurring concepts.

Topics joined in later stages of activity. The seed event that defines a topic sometimes occurs in later stages of the topic activity sequence. Previous activities within that category are then considered on-topic, though their primary events may occur outside of the corpus sampling window. References to those events tend to become ossified into shorthand tags that are referenced by corpus stories. Without stories elaborating the original events, it is hard to dereference these shorthand tags.

Sparseness of topical allusions. Some very broad topics may be signaled through a variety of disparate phrasings (e.g., "meltdown of Russian economy," "Russian financial crisis," "steep recessions in Asia and Russia," "a season of crashing banks, plunging rubles, bouncing paychecks, failing crops and rotating governments"). This is a reformulation of the well-known data sparseness problem that plagues NLP research. Some shorthand tags also evolve over time (e.g., "tropical depression," "tropical storm Mitch," "Hurricane Mitch").

Outlier and peripheral events. Many topics spawn odd outlier stories that are explicitly related to a primary topic only by short sentence linkages. The bulk of the text itself is idiosyncratic with respect to the rest of the topical text.

Some of these error classes are clearly non-tractable under current evaluation conditions, including annotation effects, lack of *a priori* topic specifications, lack of deep semantic knowledge, and violations of the single topic per story assumption, implying that an artificial operational limit will always exist for the NED track. Others are tractable through various techniques, including recourse to event level discourse analysis to detect explicit sentence linkages, vocabulary normalization within topic classes or marginalized subjects, and so forth.

Unfortunately, at least with respect to event level granularity techniques, such approaches tend to diminish evaluation performance owing to violations of evaluation biases, diminishing the incentive to tackle these issues and further exaggerating the artificial limit to NED system performance. Further, if this analysis is indeed correct, we would expect that the NED evaluation metrics should become less reliable somewhere in the 0.35 CFSD range, such that systems could be performing significantly better in "real" terms than as indicated by the official evaluation score. (We'll be happy enough to cross that bridge when we have a system operating within that range.)

3. TDT5 CORPUS

The TDT5 corpus presented some challenges during this evaluation. Weighing in at approximately 280,000 stories, it was an order of magnitude larger than either of the TDT3 or TDT4 corpora. Unfortunately, this had the effect of partly reducing our participation to an exercise in scalability. We had to substantively reengineer and optimize our classifiers for operation over the corpus. We were also forced to abandon any classification techniques that relied on part-of-speech tagging, simply because our tagger was extremely inefficient and sufficient time to convert to a new tagger was not available. (At one month for the evaluation, we had approximately 9 seconds of total processing time per story for data preprocessing and the operation of all classifiers; the part-of-speech tagger alone was taking an average of 5-20 seconds per story.) For this reason, our classifier committees were somewhat impoverished relative to TDT-2003.

We expected all NED systems to perform less accurately due to the annotation conditions this year, which favored many more topics with decreased time per topic for annotation [3]. Since the topic annotations were less likely to be complete, a larger number of misses should have been generated.

We also noticed an interesting attribute of the corpus from some of the statistics generated by our classifiers. Figure 1 lists the most frequent non-stopped, capitalized stems from the corpus, along with the percentage of stories in which the stem appears. Over 46% of all story topics reference Iraq, suggesting a certain ubiquity of that country in the news over the corpus sampling window. We might thus expect stories across topics to conflate because of an amplified presence of Iraq related verbiage across all stories generally. This would also have the effect of amping up the miss rate for classifiers employing text similarity measures.

Stem	DF (%)	Stem	DF (%)
unit	48.52	nation	27.58
iraq	46.06	iraqi	25.02
presid	36.99	china	23.94
minist	36.40	american	23.88
u.s.	34.95	washington	20.53

Figure 1. Most frequent capitalized, non-stopped stems in TDT5.

4. NED CLASSIFIERS

In this section we will detail the three classifiers that were included in our experimental runs for the NED track of TDT-2004. Our primary submission consisted of a committee composed of the Baseline classifier and two classifiers designed to shore up various errors made by the baseline technique:

1. Vector Cosine (Baseline) – A full-text similarity classifier in which each document is reduced to a Term Frequency / Inverse Document Frequency (TF/IDF) weighted feature vector of stemmed and stopped words. Vector cosine distance is used to gauge story similarity.
2. Sentence Linkage – A classifier designed to detect linking sentences in a new story that reference events in previous or future stories. The presence of such a linking sentence indicates that the stories are topically related by interest in the referenced event.
3. Location Association – A variation on the baseline approach that looks for pairs of strongly associated location entities and non-location words in a story. The constituent words in these pairs are removed from the story's vector representation and are replaced with a paired feature.

Ideally each classifier should use a judgment criterion that is orthogonal to those of other methods to increase the probability of a fruitful combination of evidence. It is not necessary that a classifier be generally effective: if it covers some classes of data better than others or if it offers increased evidence for or against a certain judgment that leads to a better overall verdict, the classifier may still be an effective component of a committee. We therefore retain classifiers with overall poor performance if they have some chance of combining well with other techniques (this is, of course, determined empirically).

Figure 2 and Figure 3 show the performance of these classifiers on the previous TDT3 and TDT4 newswire corpora (i.e., the AP and New York Times sources only), respectively. We focus on newswire sources since all relevant sources from TDT5 for the NED task this year were written-text sources.

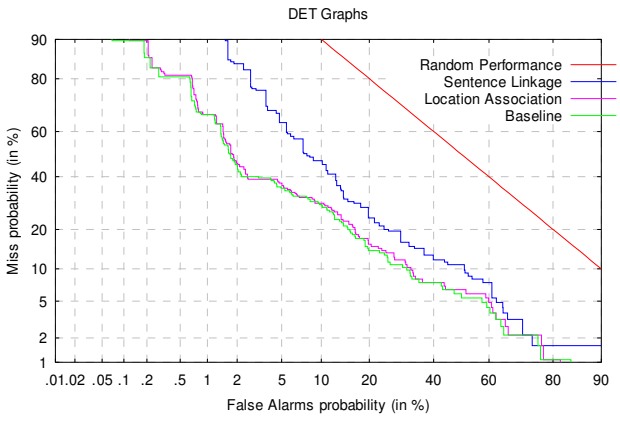

Figure 2. TDT3 newswire performance for all classifiers.

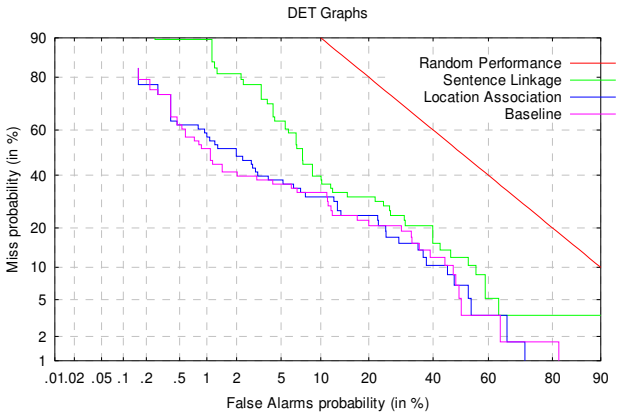

Figure 3. TDT4 newswire performance for all classifiers.

4.1. The Baseline

Our Baseline classifier is called Vector Cosine and uses the full-text similarity methodology developed by early TDT systems (see, for example, [6]). Each story is tokenized, tokens are stemmed and stopped, and a bag-of-words feature vector is constructed to represent the story. Each vector element is weighted according to its TF/IDF value, where TF and IDF are:

$$\mathrm{tf} = \log_2(\mathrm{raw\_tf} + 1)$$

$$\mathrm{idf} = \log_2((N+1)/(\mathrm{raw\_df} + 0.5))$$

The `raw_tf` and `raw_df` statistics are generated from all manual stories of TDT3 (both newswire and non-newswire). We do not employ incremental updating of frequency statistics during corpus processing. We conducted some informal experiments to evaluate the efficacy of incremental frequency updating. We found that static TDT3 manual statistics outperformed incremental updated statistics when deriving the statistics from scratch given a new corpus (e.g., the TDT4 newswire corpus). When starting from the TDT3 statistics and incrementally updating them during new corpus processing, we found performance to be about the same as for the static statistics. Given the paltry gains from incremental updating and the additional overhead associated with gathering these statistics, we opted to use the static data for the TDT5 corpus.

To evaluate the novelty of a new story, the cosine distances between its vector and all previous story vectors are determined; if any such distance is less than a threshold, the new story is deemed to be non-novel. The only tuning parameter for this classifier is the cosine similarity threshold. Figure 4 shows the optimal story- and topic-weighted CFSD values for the Baseline classifier over previous newswire corpora.

Corpus	Threshold	Story CFSD	Topic CFSD
TDT3 Newswire	0.2069	0.6253	0.5117
TDT4 Newswire	0.1907	0.5345	0.4546

Figure 4. Baseline performance.

4.2. Sentence Linkage

News stories are reports about events that occur in the world, plus contextualization to help a reader understand those events and their import. TDT topics are defined by the occurrence of a triggering, seed event in the world; antecedent and consequent events are further added to an evolving topic. We believe a critical mode of analysis, largely ignored by previous (full-text) TDT systems, is the identification of references to these topical events within a story. Correlating these event references across stories should thus aid in identifying related stories with respect to an evolving topic.

A potentially important class of event reference is what we call *linking event references*. A linking event reference is a brief description of a primary event in a new story that explicitly evokes the larger topic under consideration. When a topic has been dormant for some period of time, new stories that emerge on the topic almost always include a linking event reference that makes the relationship of the story to previous stories explicit for the audience. As we suggested in our work for TDT-2003 [2], these linking references almost always contain an explicit temporal reference (and usually a spatial reference, as well) that anchors the event geospatially. For example, a story about the Pope encouraging Catholics to donate to a relief fund after a hurricane may have only a single sentence (the linking event reference, almost certainly containing a temporal indicator) about the actual hurricane, while the rest of the text may be quite novel in terms of word similarity to previous hurricane stories; this sentence makes explicit the event that contextualizes the story. Recognizing such an event reference can thus be quite important in assigning the story to an existing hurricane topic.

A useful reference-discovery heuristic deployed in our event linking classifiers of TDT-2003 was to locate sentences that contain a temporal reference, and to hypothesize that these (and some number of adjacent sentences) constituted the event references in a story. A classifier could then search previous or future stories for additional similar event references or for whole stories that matched the event features.

For TDT-2004, we employ a variation on this theme. Rather than using temporal sentences and adjacent sentences as event reference hypotheses, we treat *every* sentence of a story as a potential coherent event reference if it meets certain characteristics of sentence length (or more accurately, unique feature length) and contains at least one capitalized feature. (This latter condition attempts to ensure that some anchoring event entity is under reference by the candidate sentence.) We then employ the following algorithm to make a novelty judgment: All event candidates in a new story are identified by determining which sentences meet the candidate criteria. Each candidate is compared to all previous stories. (All sentences and stories are stemmed, stopped, and reduced to a bag of non-weighted features prior to comparison.) If some parameterized percentage of *capitalized* features of the candidate also occur in the story and if a second threshold percentage of *all* features in the candidate is also found in the story, then the sentence is assumed to represent a reference to the previous story or to some event mentioned in that story and a link is made between the two stories. (Note that if a candidate sentence were to match only a single event reference in a previous story, it would also match the whole story since the story is a superset of all constituent sentence features.) It is assumed that the new event candidate is thus referring to an event that occurred in the past (as reported in a previous story).

We also check for the occurrence of future references by a similar means. All of the event reference candidates of a new story are saved in a list for comparison against all future stories. If the same matching criteria are met between a candidate in this list and a new story, a link is made between the source story for the matching previous candidate and the new story. It is assumed that the matching new story thus describes events that were predicted to occur by past stories.

To summarize then, a new story is considered not a first story if any of its event reference candidates (sentences that meet a particular screening process) match a previous story or if any previous reference candidates match the new story. The confidence level is the threshold percentage of all features that matched across the candidate and the linked story.

This technique seeks to address several of the error classes described in Section 2, including detection of multiple topics within a story, the insistence of a common event in high entity overlapping stories (by requiring that non-entity verbal or adjectival features of a reference be present), linking of late-stage topics through early-stage event references, and outlier and peripheral story detection through linking reference resolution.

An obvious problem with this technique (and a subject of our current research) is that it potentially matches the non-topical, ancillary event references in a story that are used to contextualize the main topics of the story. There are at several different types of such event references that must be filtered out. Many stories frequently mention similar instances of the same class of topic under discussion. For example, stories about Hurricane Mitch will frequently reference other hurricanes to provide comparative data about the impact of the hurricane; although such references are relevant to a topic at the abstract level of the topic class (hurricanes in general), it is not relevant to the specific instance topic under discussion except in its capacity as contextualization.

Another type of ancillary contextual reference includes those that draw causal connections between the topical events of a story and other non-topical events. For example, stories about mining accidents in the Ukraine may reference the general economic crisis in Russia as a means of explaining why the accident occurred. These broader contextual events tend to be considered outside the scope of the specific evaluation topic assumed to characterize the story. These types of events may be difficult to filter since some antecedent and consequent events may be considered on topic (according to the topic category descriptions) while others are not.

Figure 5 shows the optimal story- and topic-weighted CFSD values for this classifier over previous newswire corpora. The threshold parameter represents the percentage of all unique features from the linking candidate that overlapped a past or future story. Several other parameters were tuned on these previous corpora. These include parameters for identifying event reference candidates (minimum number of unique stemmed,

stopped features = 15, minimum number of capitalized features = 1) and for determining a story match (percentage of capitalized features that must overlap = 100%). Other operational parameters indicate whether days of the week should be dropped from capitalized features (no), whether the first word of a sentence should be dropped from inclusion in the capitalized feature list if it isn't already stopped out (no), use of temporal sentences or all sentences (all), and number of adjacent additional sentences to add to a sentence candidate (0).

Corpus	Threshold	Story CFSD	Topic CFSD
TDT3 Newswire	0.6500	0.7586	0.8751
TDT4 Newswire	0.4736	0.8579	0.8296

Figure 5. Sentence Linkage performance.

4.3. Location Association

One of the observations made in [7] was that two stories discussing the same type of event can be incorrectly conflated due to overlapping words that are commonly used to describe occurrences of the event type. For example, stories about plane crashes tend to contain words such as "plane", "survivor", and "accident" which are significant when compared to the corpus as a whole but not with respect to a particular plane crash. This classifier seeks to leverage the LDC definition of an event [3] to disambiguate common event words by attaching them to the geographical location where the event takes place.

A story is split into a list of location entities extracted using BBN's IdentiFinder software [1] and a list of all stemmed, stopped, non-location words. A 2x2 contingency table (Figure 6) is then constructed for each pairing of location entity (l) and non-location word (nl) in the lists.

	Number of stories in the window containing the non-location word *nl*	Number of stories in the window not containing the non-location word *nl*
Number of stories in the window containing the location *l*	A	B
Number of stories in the window not containing the location *l*	C	D

Figure 6. 2x2 contingency table.

Co-occurrence counts are captured for all stories in a sliding window that encompasses both a 10 story file (inclusive) look-ahead as well as the 20 previously seen story files. By using the look-ahead window, strongly associated pairs can be found before a judgment needs to be made for a particular story. The underlying document frequency contributions made by a story are removed when the file it belongs to moves outside of the window.

The classifier uses a vector based approach identical to the Baseline algorithm discussed in Section 4.1. In this approach, however, strongly associated pairs found in the story's text are added as new features to its vector, as follows: To prevent spurious pairings from being generated, the document frequency of the location entity and non-location word must exceed a minimum threshold of 5. The Dice coefficient is then used to measure the association strength between *l* and *nl* for each pairing in the story:

$$assoc = \frac{2A}{(A+C)+(A+B)}$$

Pairings with an `assoc` value greater than 0.7 are deemed interesting and are added into the story's vector representation. The individual words that appear in an interesting `<l, nl>` pair are then removed from the vector. The weight assigned to the pair is the sum of the non-location word's TF/IDF weight and the maximum TF/IDF weight of the words in the location entity.

Figure 7 shows the optimal story- and topic-weighted CFSD values for this classifier over previous newswire corpora. It appears that adding pairs as new features degrades the overall CFSD score relative to the Baseline. The relationship between the number of pairs added as new features and the CFSD score is unclear. There were 6 pairs added to each story on average in the TDT3 newswire set yet the results essentially mirror those of the Baseline technique. There were 8 pairs added on average to each story in the TDT4 newswire set and the results are noticeably worse. The results on TDT5 are close to the Baseline despite the fact that on average 39 pairs were added to each story.

Corpus	Threshold	Story CFSD	Topic CFSD
TDT3 Newswire	0.2074	0.6408	0.5165
TDT4 Newswire	0.1998	0.5676	0.5270

Figure 7. Location Association performance.

5. EVIDENCE COMBINATION

Once a committee of independent classifiers (each making a separate NED story novelty judgment) has been assembled, their results must be combined in a systematic manner. Our TDT-2004 submissions utilized two different evidence combination techniques: a majority voting scheme comprising all three classifiers and an authority voting scheme comprising the baseline and sentence linkage classifiers.

5.1. Majority Voting

With majority voting schemes, the members of the committee are each polled for a NED judgment and the majority decision is taken as the system decision. The intuition is that the more independent perspectives that are in agreement on a judgment for a story, the more likely that judgment is to be correct.

All of the three classifiers were tuned so as to minimize their topic-weighted CFSD score on each of the TDT3 and TDT4

newswire-only corpora. The threshold parameter to be used by each classifier on the TDT5 corpus was then estimated by taking the harmonic mean of the two thresholds learned from those corpora.

The confidence generated by the system is the average normalized distance between each majority classifier's independent confidence value and its decision threshold (dissenting classifiers thus contribute nothing to the final confidence). In the case of ties, the maximal average normalized difference between the first story versus the non-first story voters decides the system.

5.2. Authority Voting

In an authority voting committee, a single classifier is specified as the primary classifier and its judgment is the default decision. The primary classifier is typically the one that is deemed the best overall performer. (Currently, we use the Baseline classifier for this.) Other classifiers are allowed to override the default judgment if their expertise allows them to say with near certainty that a story is not novel. The intuition here is that we use the best performer of the bunch unless another classifier is extremely sure of its answer, in which case it is allowed to override the default decision and correct some of its presumed errors. We only allow not-novel overrides because of the asymmetry inherent in topic novelty determination: it is generally possible to make an instant determination of non-novelty given a new story based on some criteria, but not possible to make an instant novelty determination (the absence of evidence is not generally useful as evidence of topical absence).

When the committee is handed a story for evaluation, all classifiers starting with the primary classifier evaluate the story. If any classifier finds the story to be non-novel, a non-first judgment is made by the committee at that classifier's confidence level. If all classifiers decide the story is novel, a first-story judgment is rendered with a confidence determined by the minimum normalized confidence distance from a classifier's threshold across all classifiers; in effect, the least certain normalized first-story confidence is used.

For TDT-2004, we employed a single authority committee consisting of the Baseline classifier coupled with a Sentence Linkage classifier; the goal of the latter was to pick up stories that contained event references from other stories that were not textually similar enough for the full-text classifier to make a non-first story judgment.

Figure 8 shows the topic-weighted CFSD for the baseline operating in isolation over the TDT3 and TDT4 newswire corpora, as well as the authority committee evaluation of the same data. The TDT3 committee shows some (admittedly minor) improvement, as hoped, with approximately a half dozen errors being corrected. The TDT4 results seem to suggest that this technique hurts performance rather than helping. However, a close examination of the "errors" made by the committee reveal precisely the phenomena detailed in Section 2. One of the errors is a mislabeled topic annotation, and most of the other errors represent secondary topic references being matched in previous stories. Only one new error was actually introduced by the committee, and a handful of other errors were corrected, but because these error classes are a byproduct of the evaluation methodology, the committee improvements are penalized.

Classifiers	TDT3 CFSD	TDT4 CFSD
Baseline only	0.5117	0.4546
Baseline + Sentence Linkage	0.4912	0.4858

Figure 8. Authority committee performance.

6. EVALUATION RESULTS

Stottler Henke's official submission performed relatively well this year, outperforming other participating NED systems by small margins with story- and topic-weighted optimized CFSD scores of 0.5672 and 0.7155, respectively. The official NED DET curves [5], including our submission SHAI1, are shown in Figure 9.

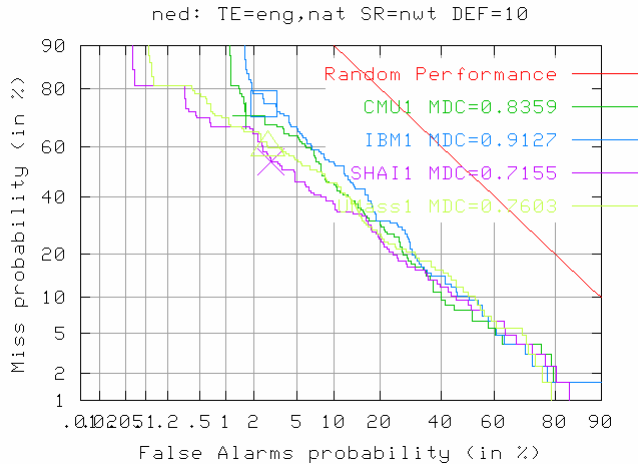

Figure 9. Official NED results for all participating sites.

The raw results are interesting in at least two ways. First, NED systems are performing worse relative to previous evaluation years. We speculated in Section 3 that this may be related to the sparser set of annotations (more stories are likely to have been missed in annotation, driving up miss error rates) and the ubiquity of news events involving Iraq (possibly causing stories to become conflated, also driving up miss error rates). Second, rather unusually, the story-weighted CFSD score is significantly less than the topic-weighted CFSD score. This is atypical relative to our experience with the TDT3 and TDT4 corpora and suggests that the TDT5 topic set has confounding properties that differ from those of previous evaluation years.

We submitted five runs for the TDT-2004 NED track, which this year consisted of the single manual condition. (The final adjudicated results can be found at [5].) Three of the runs (SHAI2, SHAI4, and SHAI5) were individual classifier runs to allow us to gauge the performance of committee constituents in

isolation (see Section 4). One run (SHAI3) was an authority committee consisting of the Baseline and Sentence Linkage classifiers (see Section 5.2). Our official submission (SHAI1) was a majority voting committee consisting of all three participating classifiers (see Section 5.1).

Figure 10 shows the committee composition, evidence combination strategy, and final optimized story- and topic-weighted CFSD scores for each of our evaluation runs. Figure 11 shows the original raw story- and topic-weighted CFSD scores given the submitted confidence thresholds.

NIST Code	Combo Method	Constituent Classifiers	Story CFSD	Topic CFSD
SHAI1	Majority	Vector Cosine Sentence Linkage Location Association	0.5672	0.7155
SHAI2	None	Vector Cosine	0.6177	0.7324
SHAI3	Authority	Vector Cosine Sentence Linkage	0.6177	0.7324
SHAI4	None	Location Association	0.6432	0.7548
SHAI5	None	Sentence Linkage	0.7767	0.8658

Figure 10. Submitted evaluation run results, optimized.

NIST Code	Constituent Classifiers	Thresh	Story CFSD	Topic CFSD
SHAI1	Vector Cosine Sentence Linkage Location Association	0.8092 0.5858 0.7964	0.8054	0.8524
SHAI2	Vector Cosine	0.8092	0.8269	0.8557
SHAI3	Vector Cosine Sentence Linkage	0.8092 0.6100	0.8341	0.8548
SHAI4	Location Association	0.7964	0.7942	0.8403
SHAI5	Sentence Linkage	0.5858	0.8042	0.9240

Figure 11. Submitted evaluation run results, original thresholds.

The authority voting committee SHAI3 showed no improvement over the unaugmented Baseline technique. This is not surprising given how far off from the optimal confidence thresholds the submitted evaluation runs appear to be (i.e., the Baseline judged stories to be non-novel far too frequently, thereby blocking any contribution by the Sentence Linkage classifier). We have had successful results combining the Sentence Linkage classifier with the Baseline only when the Baseline is run at a threshold near its optimum while the Sentence Linkage classifier is run at a threshold tighter than its optimized value. We intend to rerun the SHAI3 condition with rethresholded classifiers once we determine what the true Baseline optimum is for the TDT5 corpus. We would hope to see minor gains under this condition (though see the discussion in Section 5.2).

We were pleased (and somewhat surprised) to see that our official majority committee submission did in fact outperform all of its constituent committee members, despite the very non-optimized thresholds used for those committee constituents. While such committees typically outperform members when all members are running at optimal thresholds, we hadn't expected this apparent level of robustness. The question of how to best *a priori* tune committee members has been an open research issue. These results suggest that the committee is not particularly sensitive to the initial thresholding of its constituents. It would appear that the composite confidence computation used for committee judgments (as detailed in Section 5.1) is an effective one.

The TDT-2004 evaluation has proven to be an illuminating experience and an excellent test-bed in which to validate some of the ideas underlying development of an operational NED system. We hope to augment the reported work with those additional classifiers that could not be readied in time due to scalability issues with the TDT5 corpus. Our future research will continue in the direction of event-level analysis, as well as new techniques to tackle the tractable error classes detailed at length in Section 2. We hope to continue participation in any future TDT evaluations to proof these and other techniques en route to a viable NED operational capability.

REFERENCES

1. Bikel, D. M., Schwartz, R., and Weischedel, R. M. "An Algorithm That Learns What's In a Name", *Machine Learning*, pp. 211-231, 1999.
2. Braun, R. K. and Kaneshiro, R. "Exploiting Topic Pragmatics for New Event Detection in TDT-2003." Topic Detection and Tracking (TDT) Workshop, Gaithersburg, Maryland, November 17-18, 2003.
3. Linguistic Data Consortium (LDC). "Annotation Task Definition for 2004: V1.2", http://www.ldc.upenn.edu/Projects/TDT5/index.html, 2004.
4. National Institute of Standards and Technology (NIST). "TDT2004 Evaluation Plan: V1.2", http://www.nist.gov/speech/tests/tdt/tdt2004/evalplan.htm, 2004.
5. National Institute of Standards and Technology (NIST). "Topic Detection and Tracking 2004 Evaluation", ftp://jaguar.ncsl.nist.gov/tdt/tdt2004/eval/tdt2004_official_results_20041124/index.htm, 2004.
6. Schultz, J. M. and Liberman, M. "Topic Detection and Tracking using idf-Weighted Cosine Coefficient", *Proceedings of the DARPA Broadcast News Workshop*, pp. 189-192, 1999.
7. Yang, Y., Zhang, J., Carbonell, J., and Jin, C. "Topic Conditioned Novelty Detection", *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp 688-693, 2002.